JOURNAL OF COMPUTER AND SYSTEM SCIENCES 55, 105–118 (1997)
ARTICLE NO. SS971503

Respect to General Loss Functions*

Kenji Yamanishi†

NEC Research Institute, Inc., 4 Independence Way, Princeton, New Jersey 08540

Received June 15, 1995; revised February 9, 1996

This paper introduces a new family of deterministic and stochastic on-line prediction algorithms which work with respect to general loss functions and analyzes their behavior in terms of expected loss bounds. The algorithms use *parametric probabilistic models* regardless of the kind of loss function used. The key ideas of the algorithms are to iteratively estimate the probabilistic model using the maximum likelihood method and then to construct an optimal prediction function that minimizes the average of the loss taken with respect to the estimated probabilistic model. A future outcome is predicted using this optimal prediction function. We analyze the algorithms in the cases where the target distribution is (1) k-dimensional parametric and k is known, (2) k-dimensional parametric but k is unknown, and (3) non-parametric. For all the cases, we derive upper bounds on the expected instantaneous or cumulative losses for the algorithms with respect to a large family of loss functions satisfying the constraint introduced by Merhav and Feder. These loss bounds show new universal relations among the expected prediction accuracy, the indexes of the loss function, the complexity of the target rule, and the number of training examples.

1. INTRODUCTION

1.1. *Basic Problem and Related Work*

In this paper we consider the on-line prediction problem, formalized as follows. At each time t, an *on-line prediction algorithm* is given an instance X_t and then predicts a real-valued outcome with $\hat{Y}_t$. After the prediction, the algorithm receives a correct value Y_t and updates itself for the next prediction. This process goes on iteratively. An on-line prediction algorithm can be modified so that it outputs a probability distribution over the range of outcomes rather than a single real value. Once a loss function is given to measure the distortion of prediction, the performance of any algorithm is measured in terms of how large the loss at any given time t, which we call the *instantaneous loss*, is, or how large the total loss over all trials, which we call the *cumulative loss*, is for the algorithm. Our concern is how to design an algorithm that achieves the least instantaneous or cumulative loss.

The on-line prediction problem has extensively been addressed in statistics, information theory, and computational learning theory (e.g., [2, 6, 7, 9, 10, 12–15, 18, 22–24, 27, 32–34, 37]). In most of previous research, however, the design and analysis of on-line prediction algorithms have been restricted in the following two respects:

(A) *Specific loss functions.* Most of the previous on-line prediction algorithms have been designed to perform well for specific loss functions, e.g., logarithmic loss [2, 7, 9, 10, 12, 14, 24, 27, 34, 37], absolute loss [6, 18], quadratic loss [18, 34, 37], discrete loss [12, 13, 22, 23], etc. However, this situation, in which the design and analysis of algorithms heavily depend on the choice of a loss function, may make the research direction in this area more complicated than necessary.

Vovk [32] developed a unifying algorithm, called the *aggregating strategy*, which is designed relative to general loss functions. The aggregating strategy uses a pool of prediction algorithms, which we call *experts* (see [6, 15, 18, 33]), and its performance has been analyzed in comparison with that for the best expert, i.e., the expert with least loss. Kivinen and Warmuth [18] showed that under the twice differentiability condition for a loss function, the aggregating strategy works well for continuous outcomes in the sense that its total loss exceeds that for the best expert at most by the amount $c \ln N$, where c is a constant and N is the number of experts. Under the same smoothness conditions as in [18], Haussler *et al.* [15] derived a tight worst-case loss bound for the aggregating strategy in the case where the prediction is continuous-valued but real outcomes are binary. Vovk [33] derived a more general condition for a loss function to ensure that the aggregating strategy works well in the above sense. However, it has still remained open how to extend the results in [15, 18, 33] to a more general family of loss functions. Merhav and Feder [24] introduced an interesting family of loss functions,

* An extended abstract of this paper appeared in *Computational Learning Theory*: *Second European Conference*, *EuroCOLT'95* [36].

† The author's current address is: System Basics Research Laboratory, C&C Research Laboratories, NEC Corporation, 1-1, Miyazaki 4-chome, Miyamae-ku, Kawasaki Kanagawa 216, Japan. E-mail: yamanisi@sbl.cl.nec.co.jp.

0022-0000/97 $25.00

which seem to cover most of useful loss functions. They derived only the worst-case loss bounds with respect to them.

(B) *Worst-case analysis.* Most previous research has focused on the *worst-case loss bound analysis*, in which one makes no assumption on the target distribution so that the derived bounds hold for an arbitrary data sequence [6, 10, 15, 18, 22–24, 32, 33]. Although such worst-case analysis is important in practice, rigorous analysis of the expected losses under some statistical assumption for the target distribution can also help us to understand the statistical behavior of a given algorithm more precisely, specifically to understand how well the algorithm works depending on whether the target distribution belongs to a specific class of distributions. The *expected loss bounds* have been derived for various kinds of algorithms, including the *Bayes algorithm* [7, 9, 12, 14], the *Gibbs algorithm* [14], the *weighted average type algorithm* [34, 37], and the *maximum likelihood prediction algorithm* [2, 9, 26, 37]. These algorithms have been analyzed with respect to the discrete loss or the logarithmic loss only.

The primary contribution of this paper is to introduce a new method for the design and analysis of on-line prediction algorithms beyond the restriction (A) and (B). The algorithms proposed in this paper have the same merit as those proposed by Vovk [32], Merhav and Feder [24], and Algoet [1] in the sense that all of them are unifying algorithms designed to work with respect to *general loss functions* rather than a specific loss function. Unlike the previous ones, however, our algorithms use *parametric probabilistic models* for both cases of deterministic and stochastic prediction, regardless of the loss function used, and our analysis is done with respect to both the *expected instantaneous* and *cumulative losses* rather than the worst-case cumulative loss.

Note that in the case of deterministic prediction, Vovk's aggregating strategy does not use any probabilistic models as we do but, instead, uses a pool of deterministic prediction functions. Thus our approach may be thought of as an alternative one to Vovk's aggregating strategy for the on-line prediction problem.

We demonstrate the effectiveness of our algorithms in terms of upper bounds on the expected losses with respect to a family of general loss functions satisfying the constraint introduced by Merhav and Feder [24]. These bounds are obtained as functions of the indexes of the loss function, the complexity of the target rule, and the number of training examples.

1.2. *Summary of Results*

We consider the following three cases for the design and analysis of on-line prediction algorithms.

Case I. The conditional probability distribution (which we call the *target rule*) $f^*(Y \mid X)$ according to which each example is independently drawn, is a k-dimensional parametric model, and k is known.

Case II. The target rule f^* is a k-dimensional parametric model, but k is unknown.

Case III. The target rule f^* is nonparametric; i.e., f^* is not written as a k-dimensional parametric model for any finite k.

For Case I, we introduce the *sequential maximum likelihood prediction algorithm*, abbreviated as SML. The key ideas in SML are as follows. Let L be a given loss function. At each time t,

(1) to use a k-dimensional parametric probabilistic model $f_{\theta, k}$ regardless of the kind of loss function, where $f_{\theta, k}$ is specified by a real-valued parameter vector θ and its dimension k;

(2) to estimate θ with $\hat{\theta}_{t-1}$ using the *maximum likelihood method* from a sequence of past $t-1$ examples;

(3) to transform the probabilistic model $f_{\hat{\theta}_{t-1}, k}$ into the optimal function $b^*(f_{\hat{\theta}_{t-1}, k})$, which attains the minimum of the expected loss with respect to L, where the expectation is taken with respect to $f_{\hat{\theta}_{t-1}, k}$ and the minimum is taken over all functions. For a loss function L, we call this technique *L-transformation.*

SML takes X_t as input and then predicts the tth outcome with $\hat{Y}_t$ defined as $\hat{Y}_t = b^*(f_{\hat{\theta}_{t-1}, k})(X_t)$. The stochastic version of SML can also be defined so that the output is a probability distribution over the range rather than a single real value. When the logarithmic loss is employed, the stochastic version of SML is equivalent to the conventional on-line maximum likelihood prediction algorithm [2, 37]. In this case the cumulative loss for the stochastic SML can be thought of as the *predictive code-length* [26–30] of the data sequence relative to the hypothesis class that SML uses. Thus this research is considered to be along the direction of Rissanen's *predictive minimum description length principle* (PMDL), [26–30], which is very closely related to the *prequential analysis* explored by Dawid [9] independently. At step (2) SML replaces the parametric family by a single predictive distribution specified by an estimate of the parameter. This method is known as the plug-in approach in Dawid's forecasting system [9].

The L-transformation technique was used in [1, 24] for empirical distributions rather than for parametric models. Specifically Algoet [1] investigated the convergence of the loss per sample size for a prediction algorithm using L-transformation under general assumptions for the target distribution, including the stationary and nonstationary cases. Unlike Algoet's work, our analysis is focused on the *rate of the convergence* of the loss per sample size

(equivalently, how large the $o(m)$-term (m is sample size) in the cumulative loss is) for the case where each example is independently identically distributed.

Letting $p^*(X, Y)$ be the target distribution according to which data are generated, we prove that the expected cumulative loss for SML for sample size m is upper bounded by

$$mH(p^* : L) + \begin{cases} C\dfrac{k^{\mu/2} m^{1-(\mu/2)}}{1-\mu/2}(1+o(1)) & (0<\mu<2), \\ (Ck \ln m)(1+o(1)) & (\mu=2), \end{cases}$$

where C and μ ($0 < C < \infty$, $\mu > 0$) denote the indexes of the loss function L, of which exact definitions are given in Section 2. The quantity $H(p^* : L)$ is the least expected loss over all functions with respect to L for the target distribution $p^*(X, Y)$, and the term $o(1)$ goes to zero as m goes to infinity. Note that $\mu \leqslant 2$ holds for most of interesting loss functions. Specifically, $\mu = 2$ for the quadratic loss, and if we make a Gaussian assumption on the target distribution, $\mu = 1$ for the absolute loss.

We further show that the expected instantaneous loss for SML converges to the least expected loss as sample size t increases, with the rate of $O(C(k/t)^{\mu/2})$.

For Case II, we introduce the *weighted-average type sequential maximum likelihood prediction algorithm*, abbreviated as WA. The key idea in WA is to make a sequential prediction of a future outcome using L-transformation of the *mixture* of probabilistic models for various k. When we use the logarithmic function, the stochastic version of WA is equivalent to the conventional weighted-average type algorithm based on the maximum likelihood estimation [34, 37]. Notice here that the method for merging several predictive distributions is known as the Bayesian approach in Dawid's statistical forecasting system [9].

We prove that the expected cumulative loss for WA is upper bounded by

$$mH(p^* : L) + C(k^*)^{\mu/2} m^{1-(\mu/2)} (\ln^{\mu/2} m)(1+o(1)),$$

where m is sample size, k^* is the dimension of the real-valued parameter vector for the target rule, C and μ are indexes of the loss function L. We see that except for the case of $\mu = 2$, the cumulative loss bound for Case II is larger than that for Case I by a logarithm factor of sample size.

For Case III, we prove that the expected cumulative loss for WA for sample size m is upper bounded by

$$mH(p^* : L) + Cm \min_k \left\{ (2D_Q(f^* \parallel f_{\tilde{\theta}, k}))^{\mu/2} + \left(\frac{\xi(k, p^*) \ln m}{m}\right)^{\mu/2} (1+o(1)) \right\},$$

where $D_Q(f^* \parallel f_{\tilde{\theta}, k})$ is the approximation error of $f_{\tilde{\theta}, k}$ to the target rule f^*, measured in terms of the Kullback–Leibler divergence. Here $f_{\tilde{\theta}, k}$ is the probabilistic model that is closest to f^* in the k-dimensional models. The quantity $\xi(k, p^*)$ may be thought of as a measure of the complexity of the k-dimensional parametric hypothesis class relative to the target distribution p^*. The quantity minimized with respect to k is determined based on the optimal balance between the approximation error of the hypothesis class to the target rule and the complexity of the hypothesis, while there is a trade-off between them. This quantity can be thought of as an analogue of the *index of resolvability* (Barron and Cover [3]). Although Barron and Cover introduced this notion in the context of density estimation, we characterize the analogue in the on-line setting where general loss functions are used.

The rest of the paper is organized as follows. Section 2 defines a formal model of on-line prediction and introduces a family of loss functions. Section 3 gives a definition of SML and its examples. Section 4 gives upper bounds on the expected instantaneous and cumulative losses for SML for Case I. Section 5 defines WA. Section 6 gives an upper bound on the expected cumulative loss for WA for Case II. Section 7 gives upper bounds on the expected instantaneous and cumulative losses for SML and WA for Case III. Section 8 makes concluding remarks.

Throughout the analysis in this paper the individuality of the loss function can be reduced to the calculation of indexes μ and C. This is the reason why we can say that we offer a unifying treatment for a family of "general" loss functions.

2. ON-LINE PREDICTION MODEL

Let $\mathscr{X}$ be a subset of $\mathbf{R}^n$ for some positive integer n, or a countable set, which we call the *domain*. Let $\mathscr{Y}$ be a finite set or a subset of $\mathbf{R}$, which we call the *range*. Let $\mathscr{D} = \mathscr{X} \times \mathscr{Y}$. We denote a random variable on $\mathscr{X}$ as X and that on $\mathscr{Y}$ as Y. We denote a sequence $D_1 \cdots D_t \in \mathscr{D}^t$ ($D_j = (X_j, Y_j)$, $j = 1, ..., t$) as D^t.

A *hypothesis class* $\mathscr{H}$, which we use for prediction, is a class of functions form $\mathscr{X}$ to $\mathscr{Y}$: $\mathscr{H} = \{f(X)\}$, or a class of conditional probability distributions: $\mathscr{H} = \{f(Y \mid X)\}$. Throughout the paper, for a class of functions $\mathscr{H} = \{f(X)\}$, we call each element $f(X)$ a *deterministic rule*, and for a class of conditional distributions $\mathscr{H} = \{f(Y \mid X)\}$, we call each element $f(Y \mid X)$ a *stochastic rule* [35].

We use the *loss function* $L: \mathscr{D} \times \mathscr{H} \to \mathbf{R}^+ \cup \{0\}$ as the penalty for guessing Y using $f \in \mathscr{H}$ for given X. For any $f \in \mathscr{H}$, for any given example $D = (X, Y) \in \mathscr{D}$, we write the loss for f with respect to D as $L(D : f)$.

First consider the case where $\mathscr{Y} \subset \mathbf{R}$, and the hypothesis class $\mathscr{H}$ is a class of functions from $\mathscr{X}$ to $\mathscr{Y}$. Then, for

example, we may use the α-*loss* L_α defined as follows: for $f \in \mathscr{H}$,

$$L_\alpha(D : f) = |Y - f(X)|^\alpha \quad (\alpha \geqslant 1).$$

For $\alpha = 1$ and $\alpha = 2$, the α-loss is called the *absolute loss* and *quadratic loss*, respectively. For another example, we may use the *discrete loss* defined as $L_{\mathrm{dis}}(D : f) = \Phi(Y - f(X))$, where $\Phi(x) = 0$ if $x = 0$ and $\Phi(x) = 1$ otherwise. Further we may consider a class of general loss functions: $L(D : f) = \Psi(|Y - f(X)|)$, where Ψ is a convex continuous function such that $\Psi(0) = 0$ and $\Psi(u) > 0$ for all $u > 0$. We call Ψ a *Young function*. See e.g., [1, 4, 24] for other loss functions.

Next consider the case where $\mathscr{Y} = \{1, 2, ..., s\}$ for some positive integer s, and the hypothesis class $\mathscr{H}$ is a class of conditional probability distributions over $\mathscr{Y}$ for given X. In this case, for example, we may use the α-*loss* L_α or the *logarithmic loss* $L_{\log}$ defined as follows: for $f \in \mathscr{H}$,

$$L_\alpha(D : f) = (1 - f(Y|X))^\alpha \quad (\alpha \geqslant 1),$$
$$L_{\log}(D : f) = -\ln f(Y|X),$$

where ln denotes the natural logarithm.

For the case where $\mathscr{Y} \subset \mathbf{R}$ and $\mathscr{H}$ is a class of conditional density functions over $\mathscr{Y}$ for given X, as with the discrete case, we define the *logarithmic loss* $L_{\log}(D : f)$ for $f \in \mathscr{H}$ by

$$L_{\log}(D : f) = -\ln f(Y|X).$$

We first give a formal definition of the on-line prediction model.

Definition 1 (On-line prediction algorithm). An *on-line deterministic prediction algorithm* $\mathscr{A}$ is an algorithm that at each time t takes the tth input X_t, and then based on a sequence of the past $t-1$ examples, D^{t-1}, outputs a single real value $\hat{Y}_t = \mathscr{A}^{(t)}(X_t) \in \mathscr{Y}$ before seeing Y_t, where $\mathscr{A}^{(t)}$ is a function from $\mathscr{X}$ to $\mathscr{Y}$ determined by $\mathscr{A}$ and D^{t-1}, which we call *the* tth *predictor for* $\mathscr{A}$. After the prediction at time t, $\mathscr{A}$ receives Y_t.

An *on-line stochastic prediction algorithm* $\mathscr{A}$ is an algorithm that has the same process as the deterministic on-line prediction algorithm, except in that at each time t, $\mathscr{A}$ outputs a probability distribution $\mathscr{A}^{(t)}(Y|X_t)$ over $\mathscr{Y}$ given X_t, determined by $\mathscr{A}$ and D^{t-1}, instead of a single real value $\hat{Y}_t$, where we call $\mathscr{A}^{(t)}$ the tth *predictor for* $\mathscr{A}$.

Definition 2 (Instantaneous and cumulative losses). For a given on-line deterministic or stochastic prediction algorithm $\mathscr{A}$, let $\mathscr{A}^{(t)}$ be the tth predictor for $\mathscr{A}$ determined by D^{t-1}. For a given loss function L, we define the tth *instantaneous loss for* $\mathscr{A}$ by $L(D_t : \mathscr{A}^{(t)})$. We define the *cumulative loss for* $\mathscr{A}$ *with respect to the sequence* $D^m = D_1 \cdots D_m$ by $\sum_{t=1}^m L(D_t : \mathscr{A}^{(t)})$. Assuming that each D is independently drawn according to the target distribution $p^*(X, Y)$ on $\mathscr{D}$, we define the tth *expected instantaneous loss for* $\mathscr{A}$ by $E^t_{p^*}[L(D_t : \mathscr{A}^{(t)})]$, where $E^t_{p^*}$ denotes the expectation taken for the generation of D^t with respect to $(p^*(X, Y))^t$. We define the *expected cumulative loss for* $\mathscr{A}$ *for sample size* m by $E^m_{p^*}[\sum_{t=1}^m L(D_t : \mathscr{A}^{(t)})]$.

For a stochastic rule f, we define $b^*(f)$ by

$$b^*(f)(X) \stackrel{\mathrm{def}}{=} \arg\min_h E_f[L(D : h)]$$
$$= \arg\min_h \int f(Y|X)[L((X, Y) : h)]\, dY,$$

in the deterministic case, and

$$b^*(f)(Y|X) \stackrel{\mathrm{def}}{=} \arg\min_h E_f[L(D : h)]$$
$$= \arg\min_h \int f(Y|X)[L((X, Y) : h)]\, dY,$$

in the stochastic case, where the minimum is taken over all functions (or stochastic rules). Hereafter, we consider only the case where there exists a unique $b^*(f)$ for any $f \in \mathscr{H}$ for a given hypothesis class. It is known that when the Young function Ψ is strictly convex, for the loss function $\Psi(|Y - h(X)|)$, there exists a unique $b^*(f)$ under certain smoothness conditions for f (see [1, 16] for details).

For stochastic rules f_1, f_2 on $\mathscr{D}$, for a probability distribution $Q(X)$ over $\mathscr{X}$, we define $\Delta_X(f_1 \parallel f_2 : L)$ and $\Delta_Q(f_1 \parallel f_2 : L)$ as follows: for fixed X,

$$\Delta_X(f_1 \parallel f_2 : L) \stackrel{\mathrm{def}}{=} E_{f_1}[L(D : b^*(f_2))] - E_{f_1}[L(D : b^*(f_1))],$$
$$\Delta_Q(f_1 \parallel f_2 : L) \stackrel{\mathrm{def}}{=} E_Q \Delta_X(f_1 \parallel f_2 : L),$$

where E_f denotes the expectation taken for the generation of Y with respect to $f(Y|X)$ for fixed X, and E_Q denotes the expectation taken for the generation of X with respect to $Q(X)$. The quantity $\Delta_X(f_1 \parallel f_2 : L)$, which was first introduced by Merhav and Feder [24], can be thought of as a generalization of the Kullback–Leibler divergence into the case where a general loss function is used. In fact, if a hypothesis is a stochastic rule $f(Y|X)$ and the logarithmic loss is used as a distortion measure, then for any stochastic rule $f(Y|X)$, we have $b^*(f) = f$, and thus $\Delta_X(f_1 \parallel f_2 : L)$ is equivalent to the conditional Kullback–Leibler divergence defined as $\int f_1(Y|X) \ln(f_1(Y|X)/f_2(Y|X))\, dY$.

We consider a family of hypothesis classes and loss functions which satisfy the following assumption.

Assumption 3. For a given class of stochastic rules: $\mathscr{H} = \{f(Y|X)\}$ and a loss function L, for any $f \in \mathscr{H}$, there exists a unique $b^*(f)$, and there exist $0 < C < \infty$ and $0 < \mu \leqslant 2$ such that for every two stochastic rules $f_1, f_2 \in \mathscr{H}$, for all X, it holds that

$$\Delta_X(f_1 \parallel f_2 : L) \leqslant C(d_{v,X}(f_1, f_2))^{\mu}, \tag{1}$$

where $d_{v,X}(f_1, f_2) \stackrel{\text{def}}{=} \int |f_1(Y|X) - f_2(Y|X)|\, dY$ for fixed X, which we call the *variation distance* between f_1 and f_2.

Note that μ and C depend on $\mathscr{H}$ as well as L. This assumption has been argued by Merhav and Feder [24]. They claim that the condition (1) in Assumption 3 is not highly restrictive [24, p. 418].

EXAMPLE 4 (Indexes of the quadratic loss: Stochastic case). Letting $\mathscr{Y} = \{0, 1\}$, consider the case where a hypothesis h is a stochastic rule and we use the *quadratic loss* L_{quad} defined as $L_{\text{quad}}(D : h) = (1 - h(Y|X))^2$ for a stochastic rule h. Then we easily verify that for any stochastic rule f, we have

$$b^*(f) = f.$$

For stochastic rules f_1 and f_2, for any X, the following inequality holds:

$$\begin{aligned}\Delta_X(f_1 \parallel f_2 : L_{\text{quad}}) &= \tfrac{1}{2} \sum_Y |f_1(Y|X) - f_2(Y|X)|^2 \\ &\leqslant \tfrac{1}{2} \left(\sum_Y |f_1(Y|X) - f_2(Y|X)| \right)^2,\end{aligned}$$

where the first equation follows from Kearns and Schapire [17]. Hence in this case $C = \frac{1}{2}$ and $\mu = 2$.

EXAMPLE 5 (Indexes of the quadratic loss: Deterministic case). Letting $\mathscr{Y} = \{Y \in \mathbf{R} : |Y| \leqslant B\}$ for some $0 < B < \infty$, consider the case where a hypothesis h is a function from $\mathscr{X}$ to $\mathscr{Y}$ and we use the *quadratic loss* L_{quad} defined as $L_{\text{quad}}(D : h) = (Y - h(X))^2$ for a real-valued function h. Let $\mathscr{H} = \{f(Y|X)\}$ be an arbitrary class of stochastic rules. Then for any stochastic rule $f \in \mathscr{H}$, we have

$$b^*(f)(X) = E_f[Y|X] = \int f(Y|X)\ Y\, dY. \tag{2}$$

It is known that in general, the following equation holds. For any X,

$$\Delta_X(f_1 \parallel f_2 : L_{\text{quad}}) = (b^*(f_1)(X) - b^*(f_2)(X))^2.$$

We use this relation to obtain the following upper bound on $\Delta_X(f_1 \parallel f_2 : L_{\text{quad}})$:

$$\begin{aligned}\Delta_X(f_1 \parallel f_2 : L_{\text{quad}}) &= (b^*(f_1)(X) - b^*(f_2)(X))^2 \\ &= (E_{f_1}[Y|X] - E_{f_2}[Y|X])^2 \\ &= \left(\int Y(f_1(Y|X) - f_2(Y|X))\, dY \right)^2 \\ &\leqslant B^2 \left(\int |f_1(Y|X) - f_2(Y|X)|\, dY \right)^2 \\ &= B^2 d_{v,X}(f_1, f_2)^2.\end{aligned}$$

Hence in this case $C = B^2$ and $\mu = 2$.

EXAMPLE 6 (Indexes of the absolute loss: Deterministic case). Letting $\mathscr{Y} = \{Y \in \mathbf{R} : |Y| \leqslant B\}$ for some $0 < B \leqslant \infty$, consider the case where a hypothesis h is a function from $\mathscr{X}$ to $\mathscr{Y}$ and we use the *absolute loss* L_{abs} defined as $L_{\text{abs}}(D : h) = |Y - h(X)|$ for a real-valued function h.

Consider a hypothesis class $\mathscr{H}$ such that for any stochastic rule $f \in \mathscr{H}$, there exists a unique function $\xi(X)$ of X such that for each X,

$$\int_{\xi(X)}^{B} f(Y|X)\, dY = \int_{-B}^{\xi(X)} f(Y|X)\, dY = \tfrac{1}{2}.$$

Then we have

$$b^*(f)(X) = \xi(X). \tag{3}$$

To see this, for fixed X, define $\phi(\eta)$ by

$$\begin{aligned}\phi(\eta) &\stackrel{\text{def}}{=} E_f[L_{\text{abs}}[L(D : \eta)]] \\ &= \int_{-B}^{B} f(Y|X)\, |Y - \eta|\, dY \\ &= \int_{\eta}^{B} f(Y|X)(Y - \eta)\, dY + \int_{-B}^{\eta} f(Y|X)(\eta - Y)\, dY.\end{aligned}$$

Then we have

$$\phi'(\eta) = -\int_{\eta}^{B} f(Y|X)\, dY + \int_{-B}^{\eta} f(Y|X)\, dY.$$

We easily see that $\phi'(\eta) = 0$ only at $\eta_0 = \xi(X)$ and $\phi''(\eta_0) > 0$. Since this holds for any X, we see that $b^*(f)(X) = \xi(X)$.

For example, let $B = \infty$ and assume that $\mathscr{H}$ is a class of Gaussian distributions with a constant variance, i.e.,

$\mathscr{H} = \{ f(Y|X) = (1/\sqrt{2\pi\sigma^2}) \exp[-(Y-h(X))^2/2\sigma^2] : \sigma$ is a positive constant, $h: \mathscr{X} \to \mathbf{R}\}$, then for any $f \in \mathscr{H}$, we have

$$b^*(f)(X) = h(X).$$

For another example, letting $0 < B < \infty$, define $\mathscr{H}$ by $\mathscr{H} \stackrel{\text{def}}{=} \{ f(Y|X) = a - a^2 |Y - h(X)|$ when $|Y-h(X)| \leqslant 1/a$; otherwise $f(Y|X) = 0$: a is a positive constant, $h: \mathscr{X} \to [-B+1/a, B-1/a]\}$. Then for any $f \in \mathscr{H}$, we have

$$b^*(f)(X) = h(X).$$

For this case, we now evaluate the quantity $\Delta_X(f_1 \| f_2 : L_{\text{abs}})$. First note that $\Delta_X(f_1 \| f_2 : L_{\text{abs}})$ is decomposed as

$$\begin{aligned}
&\Delta_X(f_1 \| f_2 : L_{\text{abs}}) \\
&\quad \leqslant |E_{f_1}[L(D : b^*(f_2))] - E_{f_2}[L(D : b^*(f_2))]| \\
&\qquad + |E_{f_2}[L(D : b^*(f_2))] - E_{f_1}[L(D : b^*(f_1))]|. \qquad (4)
\end{aligned}$$

It is easy to see that the first term of the right-hand side is bounded by $(2B - 1/a) d_{v,X}(f_1, f_2)$. Below we upper bound the term $|E_{f_2}[L(D : b^*(f_2))] - E_{f_1}[L(D : b^*(f_1))]|$ in (4). Let $f_i(Y|X) = a - a^2 |Y - h_i(X)|$ $(i = 1, 2)$. Then we have

$$\begin{aligned}
&|E_{f_2}[L(D : b^*(f_2))] - E_{f_1}[L(D : b^*(f_1))]| \\
&\quad = \left| \int_{-B}^{B} (|Y - h_2(X)| f_2(Y|X) \right. \\
&\qquad \left. - |Y - h_1(X)| f_1(Y|X)) \, dY \right| \\
&\quad \leqslant \frac{1}{a^2} \int_{-B}^{B} |f_1(Y|X)(a - f_1(Y|X)) \\
&\qquad - f_2(Y|X)(a - f_2(Y|X))| \, dY. \qquad (5)
\end{aligned}$$

Now let $\Phi(f) \stackrel{\text{def}}{=} f(a-f)$. Then for any $0 \leqslant f, g \leqslant a$, we have

$$\begin{aligned}
|\Phi(f) - \Phi(g)| &\leqslant \max_{0 \leqslant \xi \leqslant a} |\Phi'(\xi)| \cdot |f - g| \\
&= a \, |f - g|.
\end{aligned}$$

Thus (5) is further upper bounded by

$$\frac{1}{a} \int_{-B}^{B} |f_1(Y|X) - f_2(Y|X)| \, dY = \frac{1}{a} d_{v,X}(f_1, f_2). \qquad (6)$$

Therefore we see from (4) that $\Delta_X(f_1 \| f_2 : L_{\text{abs}})$ is upper bounded as

$$\begin{aligned}
\Delta_X(f_1 \| f_2 : L_{\text{abs}}) &\leqslant (2B - 1/a) d_{v,X}(f_1, f_2) + \frac{1}{a} d_{v,X}(f_1, f_2) \\
&= 2B \, d_{v,X}(f_1, f_2). \qquad (7)
\end{aligned}$$

Hence in this case $C = 2B$ and $\mu = 1$.

EXAMPLE 7 (Indexes of the α-loss: Stochastic case). Letting $\mathscr{Y} = \{0, 1\}$, consider the case where a hypothesis is a stochastic rule and we use the family of α-loss functions defined as $L_\alpha(D : h) = (1 - h(Y|X))^\alpha$ for a stochastic rule h and for $\alpha > 1$.

For stochastic rules $f(Y|X)$ and $h(Y|X)$, we denote $f(0|X)$ as $f(X)$ and $h(0|X)$ as $h(X)$. Then for fixed X,

$$E_f[L_\alpha(D : h)] = f(X)(1 - h(X))^\alpha + (1 - f(X)) h^\alpha(X).$$

Thus we see that the minimum of $E_f[L_\alpha(D : h)]$ with respect to h is attained by

$$b^*(f)(X) = \frac{f^{1/(\alpha-1)}(X)}{f^{1/(\alpha-1)}(X) + (1 - f(X))^{1/(\alpha-1)}}. \qquad (8)$$

As with Example 6, let us evaluate the quantity $|E_{f_2}[L(D : b^*(f_2))] - E_{f_1}[L(D : b^*(f_1))|$. First note that

$$\begin{aligned}
&|E_{f_2}[L(D : b^*(f_2))] - E_{f_1}[L(D : b^*(f_1))]| \\
&\quad \leqslant \sum_Y \left| \frac{f_2(Y|X)(1 - f_2(Y|X))^{\alpha/(\alpha-1)}}{(f_2(Y|X)^{1/(\alpha-1)} + (1 - f_2(Y|X))^{1/(\alpha-1)})^\alpha} \right. \\
&\qquad \left. - \frac{f_1(Y|X)(1 - f_1(Y|X))^{\alpha/(\alpha-1)}}{(f_1(Y|X)^{1/(\alpha-1)} + (1 - f_1(Y|X))^{1/(\alpha-1)})^\alpha} \right|.
\end{aligned}$$

Now let $\Phi(f) \stackrel{\text{def}}{=} f(1-f)^{\alpha/(\alpha-1)} / (f^{1/(\alpha-1)} + (1-f)^{1/(\alpha-1)})^\alpha$. Since $\Phi'(f) < \infty$ for all $0 \leqslant f \leqslant 1$ and for all $\alpha > 1$, we see that there exists $0 < C < \infty$ such that

$$\begin{aligned}
|\Phi(f) - \Phi(g)| &\leqslant \max_{0 \leqslant \xi \leqslant 1} |\Phi'(\xi)| \cdot |f - g| \\
&\leqslant C \, |f - g|.
\end{aligned}$$

Thus we see that for some $0 < C < \infty$, the following inequality holds:

$$E_{f_2}[L(D : b^*(f_2))] - E_{f_1}[L(D : b^*(f_1))] \leqslant C d_{v,X}(f_1, f_2). \qquad (9)$$

Hence as with Example 6, we see that for $\alpha > 1$, for $\mu = 1$, there exists $0 < C < \infty$ satisfying (1). Note that this bound on μ is not tight for all α. Actually, $\mu = 2$ for $\alpha = 2$, as we have already seen in Example 4.

3. THE SEQUENTIAL MAXIMUM LIKELIHOOD PREDICTION ALGORITHM

We introduce here the sequential maximum likelihood prediction algorithm.

DEFINITION 8 (The sequential maximum likelihood prediction algorithm). For a fixed positive integer k, let $\mathscr{H}^{(k)} = \{ f_{\theta,k}(Y \mid X) : \theta \in \Theta(k) \subset \mathbf{R}^k \}$ be a hypothesis class of k-dimensional parametric stochastic rules, where $f_{\theta,k}$ is a stochastic rode specified by a k-dimensional real-valued parameter vector θ belonging to a set $\Theta(k) \subset \mathbf{R}^k$. Let $\hat{\theta}_{t-1}$ be the maximum likelihood estimator of θ from D^{t-1}, i.e., $\hat{\theta}_{t-1} = \arg\max_{\theta \in \Theta(k)} \prod_{j=1}^{t-1} f_{\theta,k}(Y_j \mid X_j)$ for $t \geqslant 2$, and $\hat{\theta}_0 = \theta_0$ is initially given. We suppose that for any $t \geqslant 2$, for any D^{t-1}, there exists the maximum likelihood estimator $\hat{\theta}_{t-1}$. For any given loss function L, we define $\hat{f}_k^{(t)}$ by $\hat{f}_k^{(t)} \stackrel{\text{def}}{=} b^*(f_{\hat{\theta}_{t-1},k})$.

Consider the case where $b^*(f)$ is a deterministic rule from $\mathscr{X}$ to $\mathscr{Y}$ for each $f \in \mathscr{H}^{(k)}$. The *deterministic sequential maximum likelihood prediction algorithm using* $\mathscr{H}^{(k)}$, abbreviated as the *deterministic* $\mathrm{SML}_k(L)$, is an algorithm that at each time t takes the tth input $X_t \in \mathscr{X}$, and then based on D^{t-1}, outputs $\hat{Y}_t = \hat{f}_k^{(t)}(X_t)$ before seeing Y_t. After the prediction, $\mathrm{SML}_k(L)$ receives the correct value Y_t.

Similarly, in the case where $b^*(f)$ is a stochastic rule for each $f \in \mathscr{H}^{(k)}$, the *stochastic sequential maximum likelihood prediction algorithm using* $\mathscr{H}^{(k)}$, abbreviated as the *stochastic* $\mathrm{SML}_k(L)$, is also defined as an algorithm that outputs a distribution $\hat{f}_k^{(t)}(Y \mid X_t)$ over $\mathscr{Y}$ at time t.

Note that $\mathrm{SML}_k(L)$ can be modified by replacing the maximum likelihood estimator with any other estimator $\hat{\theta}_{t-1}$ such that the central limit theorem holds, i.e., for any θ, the quantity $\sqrt{t}(\hat{\theta} - \theta)$ is asymptotically Gaussian-distributed with mean zero.

In most nontrivial cases, it would be hard to exactly compute the maximum likelihood estimator $\hat{\theta}_{t-1}$ in implementing $\mathrm{SML}_k(L)$, and it would be rather required to approximately compute it. In this paper, however, we ignore this computational problem and analyze $\mathrm{SML}_k(L)$ under the assumption that the exact value of the maximum likelihood estimator is obtained. The issue of approximating $\hat{\theta}_{t-1}$ with its applications to the actual implementation of $\mathrm{SML}_k(L)$ will be discussed elsewhere.

The cumulative loss for $\mathrm{SML}_k(L)$ is written as

$$\sum_{t=1}^{n} L(D_t : b^*(f_{\hat{\theta}_{t-1},k})). \tag{10}$$

In the specific case where $b^*(f_{\hat{\theta}_{t-1},k})$ is a stochastic rule and L is the logarithmic loss, (10) can be written as

$$-\sum_{t=1}^{n} \ln f_{\hat{\theta}_{t-1},k}(Y_t \mid X_t),$$

which can be thought of as the *predictive code-length* [26–30] of $Y_1 \cdots Y_n$ for given $X_1 \cdots X_n$, where each Y_t is sequentially encoded into a Shannon codeword with code-length (measured by nats) $-\ln f_{\hat{\theta}_{t-1},k}(Y_t \mid X_t)$, based on D^{t-1} and X_t. In this case the design of an on-line prediction algorithm that achieves the least predictive code-length can be reduced to Rissanen's *predictive minimum description length principle* (PMDL) [29, 30]. The strategy of minimizing (10) can be thought of as a general extension of PMDL to the case where either a class of deterministic or stochastic rules can be used as a hypothesis class and a general loss function is used as a distortion measure.

EXAMPLE 9 (Deterministic $\mathrm{SML}_k(L)$ w.r.t. the quadratic and absolute losses). Let $\mathscr{Y} = [-B, B] \subset \mathbf{R}$ for $B > 0$ and let a hypothesis class be a class of k-dimensional parametric stochastic rules: $\mathscr{H}^{(k)} = \{ f_{\theta,k}(Y \mid X) : \theta \in \Theta(k) \subset \mathbf{R}^k \}$. Let $\hat{\theta}_{t-1}$ be the maximum likelihood estimator of θ from D^{t-1}. Let us first employ the quadratic loss function L_{quad}. We see from (2) that at time t, the deterministic $\mathrm{SML}_k(L_{\mathrm{quad}})$ outputs

$$\begin{aligned} \hat{Y}_t = b^*(f_{\hat{\theta}_{t-1},k})(X_t) &= E_{f_{\hat{\theta}_{t-1},k}}[Y \mid X_t] \\ &= \int Y f_{\hat{\theta}_{t-1},k}(Y \mid X_t)\, dY. \end{aligned}$$

Let us next employ the absolute loss L_{abs}. Under the assumption for $\mathscr{H}$ as in Example 6, we see from (3) that at time t, the deterministic $\mathrm{SML}_k(L_{\mathrm{abs}})$ outputs

$$\hat{Y}_t = b^*(f_{\hat{\theta}_{t-1},k})(X_t) = \hat{\xi}(X_t),$$

where $\hat{\xi}(X_t)$ is defined by $\int_{\hat{\xi}(X_t)}^{B} f(Y \mid X_t)\, dY = \frac{1}{2}$.

Consider the special case where $B = \infty$ and $f_{\theta,k}(Y \mid X)$ is written as

$$f_{\theta,k}(Y \mid X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y - h_{\theta,k}(X))^2}{2\sigma^2} \right],$$

where σ is a positive constant. Here we assume that $h_{\theta,k}(X)$ can be represented as a k-dimensional parametric real-valued function, for example, as a feedforward neural network with d nodes,

$$h_{\theta,k}(X) = \sum_{j=1}^{d} c_j \phi(\mathbf{a}_j X + b_j) + c_0,$$

where $\mathbf{a}_j = (a_{j1}, ..., a_{jn}) \in \mathbf{R}^n$, $b_j, c_j, c_0 \in \mathbf{R}$ $(j = 1, ..., d)$, and ϕ is a sigmoidal function such that $\phi(z) = 1$ as $z \to \infty$ and $\phi(z) = -1$ as $z \to -\infty$. We let $\theta = (\mathbf{a}_j, ..., \mathbf{a}_d, b_1, ..., b_d, c_1, ..., c_d, c_0)$ and $k = nd + 2d + 1$.

In this case, at time t, the maximum likelihood estimator from D^{t-1} is obtained as $\hat{\theta}_{t-1} = \arg\min_\theta \sum_{j=1}^{t-1} (Y_j - h_{\theta,k}(X_j))^2$, ignoring computational complexity. We see that for the quadratic loss, the deterministic $\mathrm{SML}_k(L_{\mathrm{quad}})$ outputs $\hat{Y}_t = h_{\hat{\theta}_{t-1},k}(X_t)$ at time t.

For the absolute loss, the deterministic $\mathrm{SML}_k(L_{\mathrm{abs}})$ outputs $\hat{Y}_t = h_{\hat{\theta}_{t-1},k}(X_t)$ at time t, since $b^*(f_{\hat{\theta}_{t-1},k}) = h_{\hat{\theta}_{t-1},k}$, as seen in Example 6.

EXAMPLE 10 (Stochastic $\mathrm{SML}_k(L)$ w.r.t. the α-loss). Let $\mathscr{Y} = \{0, 1\}$ and let a hypothesis class be a class of k-dimensional parametric stochastic rules: $\mathscr{H}^{(k)} = \{f_{\theta,k}(Y \mid X) : \theta \in \Theta(k) \subset \mathbf{R}^k\}$. Let $\hat{\theta}_{t-1}$ be the maximum likelihood estimator of θ from D^{t-1}.

If we use the α-loss ($\alpha > 1$), we see from (8) that at each time t, the stochastic $\mathrm{SML}_k(L_\alpha)$ outputs $\hat{f}^{(t)}$ (we denote $\hat{f}^{(t)}(0 \mid X)$ as $\hat{f}^{(t)}(X)$) calculated as

$$\hat{f}^{(t)}(X) = \frac{f^{1/(\alpha-1)}_{\hat{\theta}_{t-1},k}(X)}{f^{1/(\alpha-1)}_{\hat{\theta}_{t-1},k}(X) + (1 - f_{\hat{\theta}_{t-1},k}(X))^{1/(\alpha-1)}}.$$

Specifically, in the case of $\alpha = 2$, $\mathrm{SML}_k(L_\alpha)$ outputs $f_{\hat{\theta}_{t-1},k}$ at each time t.

4. EXPECTED LOSS BOUNDS: k KNOWN CASE

Let us suppose that each example $D = (X, Y)$ is independently generated according to a distribution $p(X, Y) = Q(X) f(Y \mid X)$ on $\mathscr{D}$, where $Q(X)$ is a distribution over $\mathscr{X}$ and $f(Y \mid X)$ is a conditional distribution over $\mathscr{Y}$ for given X. We call $p(X, Y)$ the *target distribution* and $f(Y \mid X)$ the *target rule*. For a given loss function L, we define the *least expected loss* (with respect to p), denoted as $H(p : L)$, by

$$H(p : L) \stackrel{\mathrm{def}}{=} E_Q E_f[L(D : b^*(f))] = \min_h E_Q E_f[L(D : h)],$$

where the minimum is taken over all functions from $\mathscr{X}$ to $\mathscr{Y}$ for the deterministic prediction case (or over all stochastic rules for the stochastic prediction case).

We first give upper bounds on the expected instantaneous and cumulative losses for $\mathrm{SML}_k(L)$ in the case where it is known that the target rule belongs to a k-dimensional parametric hypothesis class.

THEOREM 11. *Let a hypothesis class be a class of k-dimensional parametric stochastic rules: $\mathscr{H}^{(k)} = \{f_{\theta,k}(Y \mid X) : \theta \in \Theta(k) \subset \mathbf{R}^k\}$. Suppose that each example is independently drawn according to $p^*(X, Y) = Q(X) f^*(Y \mid X)$ and that $f^* = f_{\theta^*,k} \in \mathscr{H}^{(k)}$. Assume that $f_{\theta,k} \in \mathscr{H}^{(k)}$ is three-times continuously differentiable with respect to θ and that for all $\theta \in \Theta(k)$, for the maximum likelihood estimator $\hat{\theta}_t$ of θ from D^t, the quantity $\sqrt{t}(\hat{\theta}_t - \theta)$ is asymptotically Gaussian-distributed*[1] *with mean zero and covariance matrix of $E_Q E_{f^*}[-\partial^2 \ln f^*/\partial\theta_i \partial\theta_j]$. Suppose that $\mathscr{H}^{(k)}$ and L satisfy Assumption 3 and let C and μ be the constants as in Assumption 3. Suppose that k is known for $f^* = f_{\theta^*,k}$. Then for any target distribution $p^*(X, Y) = Q(X) f^*(Y \mid X)$, the tth expected instantaneous loss for the deterministic and stochastic $\mathrm{SML}_k(L)$ is upper bounded as*

$$E^t_{p^*}[L(D_t : \mathrm{SML}^{(t)}_k(L))] \leqslant H(p^* : L) + C\left(\frac{k}{t}\right)^{\mu/2}(1 + o(1)), \tag{11}$$

where $\mathrm{SML}^{(t)}_k(L)$ is the tth predictor for $\mathrm{SML}_k(L)$ based on D^{t-1}, and $o(1)$ goes to zero as t goes to infinity. The expected cumulative loss for the deterministic and stochastic $\mathrm{SML}_k(L)$ for sample size m is upper bounded as

$$E^m_{p^*}\left[\sum_{t=1}^m L(D_t : \mathrm{SML}^{(t)}_k(L))\right] \leqslant mH(p^* : L) + 2^{\mu/2} C\psi(m, k, \mu), \tag{12}$$

where $\psi(m, k, \mu)$ is defined as

$$\psi(m, k, \mu) = \begin{cases} \left(\dfrac{k}{2}\right)^{\mu/2} \dfrac{m^{1-(\mu/2)}}{1 - \mu/2}(1 + o(1)) & (0 < \mu < 2), \\ ((k \ln m)/2)(1 + o(1)) & (\mu = 2). \end{cases} \tag{13}$$

Note that the $o(1)$ terms in (11) and (13) depend on k and the property of the parameterization for the hypothesis class, and thus they go to zero for *fixed* k as sample size tends to infinity.

Bound (11) shows that the expected instantaneous loss for $\mathrm{SML}_k(L)$ converges to $H(p^* : L)$, the least expected loss with respect to the target distribution, with the rate of $O(C(k/t)^{\mu/2})$ as t increases. Notice that (11) and (12) are universal in the sense that the rates of convergence depend on the loss function through μ and C only. We see that the larger μ is and the smaller C and k are, the faster the expected instantaneous loss for $\mathrm{SML}_k(L)$ converges to the least expected loss. Bounds (11) and (12) show a general relationship among the prediction accuracy, k (the complexity of the target rule), C, μ (the indexes of the loss function), and sample size.

EXAMPLE 12 (Expected cumulative loss for $\mathrm{SML}_k(L)$ w.r.t. the quadratic loss). Consider on-line deterministic prediction with respect to the quadratic loss L_{quad}. Let

[1] See Yamanishi [37, pp. 44–45, Assumption 7] see also LeCam [20, 21] and Takeuchi [31, p. 159]) for an explicit form of the sufficient condition for this central limit theorem property.

$\mathscr{Y} = \{ Y \in \mathbf{R} : |Y| \leqslant B \}$ for some $0 < B < \infty$ and $\mathscr{H}^{(k)}$ be a class of k-dimensional parametric stochastic rules. Since $C = B^2$ and $\mu = 2$ as seen in Example 5, we see from Theorem 11 that under the assumptions for $\mathscr{H}^{(k)}$ as in Theorem 11, the expected cumulative quadratic loss for the deterministic $\mathrm{SML}_k(L_{\mathrm{quad}})$ satisfies:

$$E_{p^*}^m \left[\sum_{t=1}^{m} L_{\mathrm{quad}}(D_t : \mathrm{SML}_k^{(t)}(L_{\mathrm{quad}})) \right] \leqslant mH(p^* : L_{\mathrm{quad}}) + (B^2 k \ln m)(1 + o(1)).$$

EXAMPLE 13 (Expected cumulative loss for $\mathrm{SML}_k(L)$ w.r.t. the absolute loss). Consider on-line deterministic prediction with respect to the absolute loss L_{abs} as in Example 6. Let $\mathscr{Y} = \{ Y \in \mathbf{R} : |Y| \leqslant B \}$ for some $0 < B < \infty$ and let $\mathscr{H}^{(k)} = \{ F_{\theta, k}(Y|X) = a - a^2 |Y - h_{\theta, k}(X)| :$ when $|Y - h_{\theta, k}(X)| \leqslant 1/a$; otherwise $f_{\theta, k}(Y|X) = 0 : a$ is a positive constant, $h_{\theta, k} : \mathscr{X} \to [-B + 1/a, B - 1/a], \ \theta \in \Theta(k) \subset \mathbf{R}^k \}$. Since $C = 2B$ and $\mu = 1$ as seen in Example 6, we see from Theorem 11 that under the assumptions for $\mathscr{H}^{(k)}$ as in Theorem 11, we have the upper bound on the expected cumulative absolute loss for the deterministic $\mathrm{SML}_k(L_{\mathrm{abs}})$,

$$E_{p^*}^m \left[\sum_{t=1}^{m} L_{\mathrm{abs}}(D_t : \mathrm{SML}_k^{(t)}(L_{\mathrm{abs}})) \right] \leqslant mH(p^* : L_{\mathrm{abs}}) + (4B\sqrt{km})(1 + o(1)).$$

We prepare Lemma 14 and Lemma 16 in order to prove Theorem 11. Lemma 14 gives a relation between the distance measure with respect to L and that with respect to the logarithmic loss.

LEMMA 14. *Suppose that $\mathscr{H}^{(k)}$ and L satisfy Assumption* 3 *and let C and μ be the constants as in Assumption* 3. *For a given on-line deterministic or stochastic prediction algorithm $\mathscr{A}$, we write the t th predictor as $\mathscr{A}^{(t)}$. For the loss function L, let $\tilde{\mathscr{A}}^{(t)}$ be a stochastic rule such that $b^*(\tilde{\mathscr{A}}^{(t)}) = \mathscr{A}^{(t)}$. Then for any on-line deterministic or stochastic prediction algorithm $\mathscr{A}$, for any target distribution $p^*(X, Y) = Q(X) f^*(Y|X)$, the following inequality holds*:

$$E_{p^*}^{t-1}[\Delta_Q(f^* \parallel \tilde{\mathscr{A}}^{(t)} : L)] \leqslant 2^{\mu/2} C (E_{p^*}^t[L_{\log}(D_t : \tilde{\mathscr{A}}^{(t)})] - H(p^* : L_{\log}))^{\mu/2}. \quad (14)$$

Proof. For stochastic rules f_1 and f_2, let us denote $D_X(f_1 \parallel f_2) \stackrel{\mathrm{def}}{=} \int f_1(Y|X) \ln(f_1(Y|X)/f_2(Y|X))\, dY$, which we call the conditional *Kullback–Leibler divergence* between f_1 and f_2 for given X. Observe first that the following equations hold:

$$\begin{aligned} &E_{p^*}^t[L_{\log}(D_t : \tilde{\mathscr{A}}^{(t)})] \\ &\quad = E_{p^*}^t[-\ln \tilde{\mathscr{A}}^{(t)}(Y_t|X_t)] \\ &\quad = E_{p^*}[-\ln f^*(Y_t|X_t)] + E_{p^*}^t\left[\ln \frac{f^*(Y_t|X_t)}{\tilde{\mathscr{A}}^{(t)}(Y_t|X_t)}\right] \\ &\quad = H(p^* : L_{\log}) + E_{p^*}^{t-1} E_Q[D_X(f^* \parallel \tilde{\mathscr{A}}^{(t)})] \\ &\quad \geqslant H(p^* : L_{\log}) + \frac{1}{2} E_{p^*}^{t-1} E_Q[(d_{v,X}(f^*, \tilde{\mathscr{A}}^{(t)}))^2]. \end{aligned} \quad (15)$$

Inequality (15) is derived using the following sublemma.

SUBLEMMA 15 (Kullback [19]). *For any two stochastic rules f_1 and f_2, for any X,*

$$\tfrac{1}{2}(d_{v,X}(f_1, f_2))^2 \leqslant D_X(f_1 \parallel f_2). \quad (16)$$

We are further able to lower-bound Eq. (15) as

$$\begin{aligned} &H(p^* : L_{\log}) + \frac{1}{2} E_{p^*}^{t-1} E_Q[(d_{v,X}(f^*, \tilde{\mathscr{A}}^{(t)}))^2]. \\ &\quad = H(p^* : L_{\log}) + \frac{1}{2} E_{p^*}^{t-1} E_Q[((d_{v,X}(f^*, \tilde{\mathscr{A}}^{(t)}))^{\mu})^{2/\mu}] \\ &\quad \geqslant H(p^* : L_{\log}) + \frac{1}{2} E_{p^*}^{t-1} E_Q\left[\left(\frac{1}{C}\Delta_X(f^* \parallel \tilde{\mathscr{A}}^{(t)} : L)\right)^{2/\mu}\right] \end{aligned} \quad (17)$$

$$\begin{aligned} &\geqslant H(p^* : L_{\log}) + \frac{1}{2C^{2/\mu}} (E_{p^*}^{t-1} E_Q[\Delta_X(f^* \parallel \tilde{\mathscr{A}}^{(t)} : L)])^{2/\mu} \\ &= H(p^* : L_{\log}) + \frac{1}{2C^{2/\mu}} (E_{p^*}^{t-1}[\Delta_Q(f^* \parallel \tilde{\mathscr{A}}^{(t)} : L)])^{2/\mu}. \end{aligned} \quad (18)$$

Inequality (17) is derived using (1) in Assumption 3. Inequality (18) is derived using Jensen's inequality with the fact that $x^{2/\mu}$ is a convex function of x for $0 < \mu \leqslant 2$. Hence we have

$$\begin{aligned} &E_{p^*}^t[L_{\log}(D_t : \tilde{\mathscr{A}}^{(t)})] \\ &\quad \geqslant H(p^* : L_{\log}) \\ &\qquad + (1/2C^{2/\mu})(E_{p^*}^{t-1}[\Delta_Q(f^* \parallel \tilde{\mathscr{A}}^{(t)} : L)])^{2/\mu}. \end{aligned}$$

This is equivalent to Eq. (14). ∎

Lemma 16 gives an expected instantaneous loss for $\mathrm{SML}_k(L_{\log})$ with respect to the logarithmic loss.

LEMMA 16 (Čencov [5], Amari and Murata [2], Yamanishi [37]). *Suppose that k is known for $f^* = f_{\theta^*, k}$. Under the assumptions for $\mathscr{H}^{(k)}$ as in Theorem* 11, *for any*

$p^*(X, Y) = Q(X)\, f^*(Y|X)$, an upper bound on the tth expected instantaneous logarithmic loss for the stochastic $\mathrm{SML}_k(L_{\log})$ *is given by*

$$E^t_{p^*}[L_{\log}(D_t : \mathrm{SML}^{(t)}_k(L_{\log}))] = H(p^* : L_{\log}) + \frac{k}{2t} + O\left(\frac{1}{t^{3/2}}\right). \tag{19}$$

Proof of Theorem 11. First observe that the following equations hold:

$$\begin{aligned} &E^t_{p^*}[L(D_t : \mathrm{SML}^{(t)}_k(L))] \\ &\quad = E^t_{p^*}[L(D_t : b^*(f^*))] + (E^t_{p^*}[L(D_t : \mathrm{SML}^{(t)}_k(L))] \\ &\qquad - E^t_{p^*}[L(D_t : b^*(f^*))]) \\ &\quad = H(p^* : L) + E^{t-1}_{p^*}[\Delta_Q(f^* \,\|\, \mathrm{SML}^{(t)}_k(L_{\log}) : L)]. \end{aligned} \tag{20}$$

Since $\mathrm{SML}^{(t)}_k(L) = b^*(\mathrm{SML}^{(t)}_k(L_{\log}))$, setting $\mathscr{A}^{(t)} = \mathrm{SML}^{(t)}_k(L)$ in (14) and then plugging it into (20) yields

$$\begin{aligned} &E^t_{p^*}[L(D_t : \mathrm{SML}^{(t)}_k(L))] \\ &\quad \leqslant H(p^* : L) + 2^{\mu/2} C(E^t_{p^*}[L_{\log}(D_t : \mathrm{SML}^{(t)}_k(L_{\log}))] \\ &\qquad - H(p^* : L_{\log}))^{\mu/2}. \end{aligned} \tag{21}$$

Plugging (19) into (21) yields (11). Since

$$E^m_{p^*}\left[\sum_{t=1}^m L(D_t : \mathrm{SML}^{(t)}_k(L))\right] = \sum_{t=1}^m E^t_{p^*}[L(D_t : \mathrm{SML}^{(t)}_k(L))]$$

by the information-theoretic chain-rule (see, e.g., Cover and Thomas [8]), summing both sides of (11) with respect to t gives

$$E^m_{p^*}\left[\sum_{t=1}^m L(D_t : \mathrm{SML}^{(t)}_k(L))\right] \leqslant mH(p^* : L) + 2^{\mu/2} C\left(\sum_{t=\ell_m}^m \left(\frac{k}{2t}\right)^{\mu/2} + o(\ell_m)\right),$$

where $\ell_m = m^{1-\mu/2}$ $(0 < \mu < 2)$ and $\ell_m = \ln m$ $(\mu = 2)$. Since

$$\sum_{t=\ell_m}^m \left(\frac{k}{2t}\right)^{\mu/2} \leqslant \int_1^m \left(\frac{k}{2x}\right)^{\mu/2} dx + o(\ell_m) = \psi(m, k, \mu)$$

as in (13), we have (12). This completes the proof of Theorem 11. ∎

Note that (11) and (12) also hold for any estimator $\hat{\theta}_{t-1}$ other than the maximum likelihood estimator whenever it satisfies the same property as (19) corresponding to the central limit theorem.

Further note that the key to deriving (11) and (12) is the reduction of the general loss case to the logarithmic loss case. Hence replacing condition (1) with a more direct relation between L and the logarithmic loss,

$$\Delta_X(f_1 \,\|\, f_2 : L) \leqslant C(D_X(f_1 \,\|\, f_2))^v \tag{22}$$

for $0 < v \leqslant 1$, we may obtain a stronger upper bound than (11), in which the term $C(k/t)^{\mu/2}$ in (11) may be replaced with $C(k/2t)^v$. Nevertheless, we offer the condition (1) and the bound of the form of (11), since in most nontrivial cases, as seen in Examples 4–6, it is easier to estimate μ in (1) than to estimate v in (22) directly without using Sublemma 15, and thus (1) might be a more useful condition than (22) for actual applications.

5. THE WEIGHTED-AVERAGE TYPE ALGORITHM

We introduce here the weighted-average type sequential maximum likelihood prediction algorithm, which uses as a hypothesis class a union of probabilistic models with various dimensions of parameter vectors. We use this algorithm when the target rule is a k-dimensional parametric model but k is unknown, or the target rule is essentially non-parametric. When k is unknown for the target rule, if we use a single k-dimensional probabilistic model with quite small k, we may not be able to model the target rule well enough to predict future data accurately. On the other hand, if we use a single k-dimensional probabilistic model with quite large k, such a model may be overfitting past examples, and thus will make the prediction accuracy worse. Our algorithm uses a number of probabilistic models with various k and takes a weighted average of the predicted probabilities over all ks. Here a weight for k is calculated so that it is proportional to the cumulative product of the maximum likelihood for the k-dimensional parametric model. The algorithm is thereby expected to asymptotically perform as well as the algorithm using a single probabilistic model with the optimal k.

DEFINITION 17 (The weighted-average type algorithm). For a given positive integer K, let a hypothesis class be $\mathscr{H} = \bigcup_{k=1}^K \mathscr{H}^{(k)}$, where $\mathscr{H}^{(k)} = \{f_{\theta,k} : \theta \in \Theta(k) \subset \mathbf{R}^k\}$ is a class of k-dimensional parametric stochastic rules. We assume that for each k, the class $\mathscr{H}^{(k)}$ satisfies the assumptions as in Definition 8 and Theorem 11. First consider the case where for a given loss function L, for each $f \in \mathscr{H}$, we are given $b^*(f)$ as a deterministic rule from $\mathscr{X}$ to $\mathscr{Y} \subset \mathbf{R}$. The *deterministic weighted-average type sequential maximum likelihood prediction algorithm using* $\mathscr{H}$, abbreviated as the

deterministic WA(L), is an algorithm that at each time t takes the tth input $X_t \in \mathcal{X}$, and then based on D^{t-1}, outputs $\hat{Y}_t = b^*(\bar{f}_t)(X_t)$ before seeing Y_t, where $\bar{f}_t(Y \mid X)$ is a conditional probability distribution over $\mathcal{Y}$ defined by

$$\bar{f}_t(Y \mid X) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \hat{w}(k \mid D^{t-1})\, f_{\hat{\theta}_{t-1},k}(Y \mid X),$$

where

$$\hat{w}(k \mid D^{t-1}) \stackrel{\text{def}}{=} \frac{\pi(k) \prod_{j=1}^{t-1} f_{\hat{\theta}_{j-1},k}(Y_j \mid X_j)}{\sum_{k'=1}^{K} \pi(k') \prod_{j=1}^{t-1} f_{\hat{\theta}_{j-1},k'}(Y_j \mid X_j)},$$

$$\hat{\theta}_{j-1} \stackrel{\text{def}}{=} \arg\max_{\theta \in \Theta(k)} \prod_{l=1}^{j-1} f_{\theta,k}(Y_l \mid X_l),$$

where $\hat{\theta}_{j-1}$ is the maximum likelihood estimator of θ from D^{j-1}, we define $\hat{w}(k \mid D^0) = \pi(k)$ as an initially given prior probability distribution over $\{1, ..., K\}$, and $\hat{\theta}_0 = \theta_0$ are initially given for each k. After the tth prediction WA(L) receives Y_t.

Similarly, in the case where $b^*(f)$ is a stochastic rule, the *stochastic weighted-average type sequential maximum likelihood prediction algorithm using* $\mathcal{H}$, which we abbreviate as the *stochastic* WA(L), is also defined as an algorithm that outputs a stochastic rule $\hat{f}_t(Y \mid X_t) = b^*(\bar{f}_t)(Y \mid X_t)$ at time t.

In Definition 17, for a finite K, a prior distribution $\pi(k)$ may be, for example, set to be $1/K$ for all k if all ks are equally probable. An alternative is to set $\pi(k) = e^{-\ell^*(k)} / \sum_{k=1}^{K} e^{-\ell^*(k)}$. Here $\ell^*(k)$ is Rissanen's integer coding length [25] defined by $\ell^*(k) = (\ln 2)(\log c + \log k + \log\log k + \cdots)$, where the sum is taken over all positive terms and c is about 2.865.

The stochastic WA(L) for the logarithmic loss is equivalent with the *weighted-average type algorithm* proposed by Yamanishi [37], which is truly different from the Bayes algorithm in that each weight for k in WA($L_{\log}$) is proportional to the cumulative product of the maximum likelihood for k, while the weight in the Bayes algorithm is proportional to the integral of the likelihood with respect to the parameter. The weighted-average type algorithm can be thought of as a hybrid of the Bayes algorithm and the sequential maximum likelihood parameter estimation method.

6. EXPECTED LOSS BOUND: k UNKNOWN CASE

The following theorem gives an upper bound on the expected cumulative loss for WA(L) in the case where k^* is unknown for the target rule $f^* = f_{\theta^*,k^*}$.

THEOREM 18. *Let a hypothesis class be* $\mathcal{H} = \bigcup_{k=1}^{K} \mathcal{H}^{(k)}$ *for given* K, *where* $\mathcal{H}^{(k)} = \{f_{\theta,k}(Y \mid X) : \theta \in \Theta(k) \subset \mathbf{R}^k\}$ *is a class of* k*-dimensional parametric stochastic rules. Suppose that each example is independently drawn according to* $p^*(X, Y) = Q(X) f^*(Y \mid X)$ *and that* $f^* = f_{\theta^*,k^*} \in \mathcal{H}^{(k^*)} \subset \mathcal{H}$, *where* k^* *is unknown. Suppose that for each* k, *the class* $\mathcal{H}^{(k)}$ *and* L *satisfy the assumptions as in Theorem* 11. *Then the expected cumulative loss for the deterministic and stochastic* WA(L) *using* $\mathcal{H}$ *for sample size* m *is upper bounded as*,

$$E_{p^*}^{m}\left[\sum_{t=1}^{m} L(D_t : \mathrm{WA}^{(t)}(L))\right] \leqslant mH(p^* : L) + C(k^*)^{\mu/2}\, m^{1-(\mu/2)} (\ln^{\mu/2} m)(1 + o(1)). \tag{23}$$

Comparing (23) with (12) we see that for $0 < \mu < 2$, the cumulative loss bound for the k^*-unknown case (k^* is the number of parameters for the target rule) is larger by a logarithmic factor of sample size than that for the k^*-known case.

Proof of Theorem 18. First note according to [34, 37] (see also [10, 32]) that the cumulative logarithmic loss for WA($L_{\log}$) is upper bounded by that for SML$_k(L_{\log})$ plus $-\ln \pi(k)$, for all k such that $1 \leqslant k \leqslant K$. That is, we have the following lemma.

LEMMA 19. *For any* $1 \leqslant k \leqslant K$, *we have*

$$\sum_{t=1}^{m} L_{\log}(D_t : \mathrm{WA}^{(t)}(L_{\log})) \leqslant \sum_{t=1}^{m} L_{\log}(D_t : \mathrm{SML}_k^{(t)}(L_{\log})) - \ln \pi(k). \tag{24}$$

Next observe that for any target distribution p^*, the following inequalities hold:

$$\begin{aligned}
&E_{p^*}^{m}\left[\sum_{t=1}^{m} L(D_t : \mathrm{WA}^{(t)}(L))\right] \\
&\leqslant mH(p^* : L) + 2^{\mu/2} C \\
&\quad \times \sum_{t=1}^{m} (E_{p^*}^{t}[L_{\log}(D_t : \mathrm{WA}^{(t)}(L_{\log}))] - H(p^* : L_{\log}))^{\mu/2} \\
&= mH(p^* : L) + 2^{\mu/2} Cm \\
&\quad \cdot \frac{1}{m} \sum_{t=1}^{m} (E_{p^*}^{t}[L_{\log}(D_t : \mathrm{WA}^{(t)}(L_{\log}))] - H(p^* : L_{\log}))^{\mu/2} \\
&\leqslant mH(p^* : L) + 2^{\mu/2} Cm \\
&\quad \times \left(\frac{1}{m} E_{p^*}^{m}\left[\sum_{t=1}^{m} L_{\log}(D_t : \mathrm{WA}^{(t)}(L_{\log}))\right] - H(p^* : L_{\log})\right)^{\mu/2}.
\end{aligned} \tag{25}$$

Here the first inequality is derived as with (21). The second inequality is derived using Jensen's inequality with the fact that $x^{\mu/2}$ is a concave function for $0<\mu\leqslant 2$.

With regard to the right-hand side of (25), plugging the true dimension k^* for the target rule into k in (24) gives the inequality:

$$\frac{1}{m}E^m_{p^*}\left[\sum_{t=1}^{m} L_{\log}(D_t : \mathrm{WA}^{(t)}(L_{\log}))\right] - H(p^* : L_{\log}) \leqslant \frac{1}{m}\left(E^m_{p^*}\left[\sum_{t=1}^{m} L_{\log}(D_t : \mathrm{SML}^{(t)}_{k^*}(L_{\log}))\right] - \ln \pi(k^*) - mH(p^* : L_{\log})\right). \quad (26)$$

Since $E^t_{p^*}[L_{\log}(D_t : \mathrm{SML}^{(t)}_{k^*}(L_{\log}))] = H(p^* : L_{\log}) + k^*/2t + O(1/t^{3/2})$ by Lemma 16, we have

$$\begin{aligned} &E^m_{p^*}\left[\sum_{t=1}^{m} L_{\log}(D_t : \mathrm{SML}^{(t)}_{k^*}(L_{\log}))\right] - \ln \pi(k^*) - mH(p^* : L_{\log}) \\ &= \sum_{t=1}^{m} E^t_{p^*}[L_{\log}(D_t : \mathrm{SML}^{(t)}_{k^*}(L_{\log}))] - \ln \pi(k^*) - mH(p^* : L_{\log}) \\ &\leqslant \int_1^m (k^*/2x)\, dx + o(\ln m) \\ &= ((k^* \ln m)/2)(1+o(1)). \end{aligned}$$

Plugging this into (26) and then combining it with (25) yields (23). This completes the proof of Theorem 18. ∎

7. EXPECTED LOSS BOUNDS: NON PARAMETRIC CASE

Next we consider the nonparametric case where the target rule does not belong to a k-dimensional parametric hypothesis class for any finite k.

THEOREM 20. *Let $\mathscr{H}^{(k)} = \{f_{\theta,k}(Y|X) : \theta \in \Theta(k) \subset \mathbf{R}^k\}$ be a class of k-dimensional parametric stochastic rules. Suppose that each example is independently drawn according to $p^*(X, Y) = Q(X) f^*(Y|X)$, where f^* does not belong to $\mathscr{H}^{(k)}$ for any finite k. Suppose that for each k, the class $\mathscr{H}^{(k)}$ and L satisfy Assumption 3 and let C and μ be the constants in Assumption 3. For each k, let $I_k \stackrel{\mathrm{def}}{=} E_{p^*}[(\partial \ln f_{\theta,k}(Y|X)/\partial\theta_i)^{\mathrm{T}}(\partial \ln f_{\theta,k}(Y|X)/\partial\theta_j)|_{\theta=\tilde{\theta}}]$ and $J_k \stackrel{\mathrm{def}}{=} E_{p^*}[-\partial^2 \ln f_{\theta,k}(Y|X)/\partial\theta_i\partial\theta_j|_{\theta=\tilde{\theta}}]$, where $\tilde{\theta}(\in \Theta(k)) = \arg\min_{\theta\in\Theta(k)} E_{p^*}[-\ln f_{\theta,k}(Y|X)]$, assuming the existence of $\tilde{\theta}$ and regularity of J_k for each k. Let*

$$\xi(k, p^*) \stackrel{\mathrm{def}}{=} \mathrm{tr}(J_k^{-1} I_k),$$

where tr *A means the trace of the matrix A. Then for any p^*, the tth expected instantaneous loss for the deterministic and stochastic $\mathrm{SML}_k(L)$ using $\mathscr{H}^{(k)}$ is upper bounded as*

$$\begin{aligned} &E^t_{p^*}[L(D_t : \mathrm{SML}^{(t)}_k(L))] \\ &\leqslant H(p^* : L) + C\left((2D_Q(f^* \,\|\, f_{\tilde{\theta},k}))^{\mu/2} + \left(\frac{\xi(k, p^*)}{t}\right)^{\mu/2}(1+o(1))\right), \end{aligned} \quad (27)$$

where $D_Q(f^* \,\|\, f_{\tilde{\theta},k}) \stackrel{\mathrm{def}}{=} E_Q[D_X(f^* \,\|\, f_{\tilde{\theta},k})]$. Let $\mathscr{H} = \bigcup_{k=1}^{K} \mathscr{H}^{(k)}$ for a given positive integer K. Then the expected cumulative loss for the deterministic and stochastic $\mathrm{WA}(L)$ *using $\mathscr{H}$ for sample size m is upper bounded as*

$$\begin{aligned} &E^m_{p^*}\left[\sum_{t=1}^{m} L(D_t : \mathrm{WA}^{(t)}(L))\right] \\ &\leqslant mH(p^* : L) + Cm \min_{1\leqslant k\leqslant K}\left\{(2D_Q(f_{\tilde{\theta},k} \,\|\, f^*))^{\mu/2} + \left(\frac{\xi(k, p^*)\ln m}{m}\right)^{\mu/2}(1+o(1))\right\}. \end{aligned} \quad (28)$$

Note that if f^* is written as a k-dimensional parametric model, i.e., $f^* = f^*_{\theta^*,k} \in \mathscr{H}^{(k)}$, then $D_Q(f^* \,\|\, f_{\tilde{\theta},k}) = 0$ and I_k is equal to J_k; therefore $\xi(k, p^*) = k$. In this case the bounds (27) and (28) coincide with (11) and (23), respectively.

The quantity $(2D_Q(f^* \,\|\, f_{\tilde{\theta},k}))^{\mu/2}$ in (27) and (28) can be thought of as an *approximation error* of the hypothesis class $\mathscr{H}^{(k)}$ to f^* measured in terms of the Kullback–Leibler divergence, also depending on the loss function through μ. That is, $(2D_Q(f^* \,\|\, f_{\tilde{\theta},k}))^{\mu/2}$ measures how close to $\mathscr{H}^{(k)}$ the target rule is. Hence (27) relates the instantaneous loss for $\mathrm{SML}_k(L)$ to the approximation error of $\mathscr{H}^{(k)}$ to the target rule, as well as the term $O(C(\xi(k, p^*)/t)^{\mu/2})$, which converges to zero as the sample size increases.

The quantity $((\xi(k, p^*)\ln m)/m)^{\mu/2}$ in (28) is related to the complexity of $\mathscr{H}^{(k)}$. Thus we see that the quantity minimized with respect to k in (28) is determined by the optimal balance between the approximation error of $\mathscr{H}^{(k)}$ to f^* and the complexity of $\mathscr{H}^{(k)}$, while there is a trade-off between them. This minimized quantity can be thought of as an analogue of Barron and Cover's *index of resolvability* [3], which was introduced to characterize the rate of convergence of the minimum complexity estimator. Through (28) the analogue introduced here also characterizes how the choices of p^*, $\mathscr{H}$, and L affect the rate of convergence of the expected cumulative loss for $\mathrm{WA}(L)$ per sample size to the least expected loss.

Proof of Theorem 20. First notice the fact from [2] that the following asymptotics holds for the expected instantaneous loss for $\mathrm{SML}_k(L_{\log})$.

LEMMA 21 (Amari and Murata [2]). *Under the same notation and assumption as Theorem* 20, *for any* $p^*(X, Y) = Q(X)\, f^*(Y|X)$, *we have the upper bound on the t th expected instantaneous logarithmic loss for the stochastic* $\mathrm{SML}_k(L_{\log})$,

$$E^t_{p^*}[L_{\log}(D_t : \mathrm{SML}^{(t)}_k(L_{\log}))] = H(p^* : L_{\log}) + D_Q(f^* \parallel f_{\tilde{\theta},k}) + \frac{\xi(k, p^*)}{2t} + o\left(\frac{1}{t}\right). \tag{29}$$

Plugging (29) into (21), we have

$$\begin{aligned} E^t_{p^*}[L(D_t : \mathrm{SML}^{(t)}_k(L))] &\leqslant H(p^* : L) \\ &\quad + C\left(2D_Q(f^* \parallel f_{\tilde{\theta},k}) + \frac{\xi(k, p^*)}{t}(1 + o(1))\right)^{\mu/2} \\ &\leqslant H(p^* : L) + C\left((2D_Q(f^* \parallel f_{\tilde{\theta},k}))^{\mu/2}\right. \\ &\quad \left. + \left(\frac{\xi(k, p^*)}{t}\right)^{\mu/2}(1 + o(1))\right). \end{aligned}$$

This yields Eq. (27). Here we have used the fact that $(X + Y)^{\mu/2} \leqslant X^{\mu/2} + Y^{\mu/2}$ for $0 < \mu \leqslant 2$ and for $X, Y \geqslant 0$, to derive the last inequality.

As with the proof of Theorem 18, summing (29) with respect to t yields

$$\begin{aligned} \frac{1}{m}\left[E^m_{p^*}\sum_{t=1}^{m} L_{\log}(D_t : \mathrm{WA}^{(t)}(L_{\log}))\right] - H(p^* : L_{\log}) \\ \leqslant D_Q(f_{\tilde{\theta},k} \parallel f^*) + \frac{\xi(k, p^*)\ln m}{2m}(1 + o(1)), \end{aligned}$$

for all $1 \leqslant k \leqslant K$. Plugging this into (25) gives

$$\begin{aligned} E^m_{p^*}\left[\sum_{t=1}^{m} L(D_t : \mathrm{WA}^{(t)}(L))\right] \\ \leqslant mH(p^* : L) + Cm\left((2D_Q(f^* \parallel f_{\tilde{\theta},k}))^{\mu/2}\right. \\ \left. + \left(\frac{\xi(k, p^*)\ln m}{m}\right)^{\mu/2}(1 + o(1))\right). \end{aligned} \tag{30}$$

Since (30) holds for any $1 \leqslant k \leqslant K$, minimizing (30) with respect to k gives (28). This completes the proof of Theorem 20. ∎

8. CONCLUDING REMARKS

This paper has introduced a new family of on-line prediction algorithms based on probabilistic modeling of data generation and the maximum likelihood method. This family is given in a quite general form, which is applicable to both of the deterministic and stochastic prediction problems with respect to most useful loss functions. Our strategy can be thought of as a probabilistic approach to the issue of on-line prediction with respect to general loss functions, while Vovk's aggregating strategy is a non probabilistic one to the same issue. For the proposed algorithms we have derived universal upper bounds on the cumulative losses for them as functions of the indexes μ and C of a loss function. Depending on the loss function and the hypothesis class, it is sometimes very easy to get tight bounds on the indexes but sometimes analytically hard. It remains for future study to develop a universal method for calculating the indexes of a loss function.

An interesting possible extension of the results in this paper would be to weaken the i.i.d. assumption for data generation by replacing it with a more natural stochastic assumption like the Markovian one. It would also be an important open issue to provide lower bounds on the losses with which we have dealt, to discuss how tight our current upper bounds are.

ACKNOWLEDGMENTS

The author thanks Leonid Gurvits for helpful discussion and Jason Catlett for reading an earlier draft of this paper. He also thanks two anonymous reviewers for their valuable comments.

REFERENCES

1. P. H. Algoet, The strong law of large numbers for sequential decisions under uncertainty, *IEEE Inform. Theory* **IT-40** (1994), 609–633.
2. S. Amari and N. Murata, Statistical theory of learning curves under entropic loss criterion, *Neural Comput.* **5** (1993), 140–153.
3. A. R. Barron and T. M. Cover, Minimum complexity density estimation, *IEEE Trans. Inform. Theory* **IT-37** (1991), 1034–1054.
4. J. O. Berger, "Statistical Decision Theory and Bayesian Analysis," Springer Verlag, New York/Berlin, 1980.
5. N. N. Čencov, Evaluation of an unknown distribution density from observations, *Soviet Math.* **3** (1962), 1559–1562.
6. N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. Schapire, and M. K. Warmuth, How to use expert advise, *in* "Proceedings of The Twenty-Fifth Annual ACM Symposium on Theory of Computing," pp. 429–438, Assoc. Comput. Mach., New York, 1993.
7. B. Clarke and A. Barron, Information-theoretic asymptotics of Bayes methods, *IEEE Trans. Inform. Theory* **IT-36** (1990), 453–471.
8. T. M. Cover and J. A. Thomas, "Elements of Information Theory," Wiley-Interscience, New York, 1991.
9. A. Dawid, Statistical theory: The prequential approach, *J. R. Stat. Soc. A* (1984), 278–292.
10. A. DeSantis, G. Markowsky, and M. N. Wegman, Learning probabilistic prediction functions, *in* "Proceedings of the First Annual Workshop on Computational Learning Theory," pp. 312–328, Morgan Kaufmann, San Mateo, CA, 1988.

11. R. A. Fisher, "Statistical Methods and Scientific Inference," Oliver & Boyd, Edingburgh, 1951.
12. D. Haussler and A. Barron, How well does the Bayes method work in on-line predictions of $\{+1, -1\}$-values?, *in* "Proceedings of the Third NEC Symposium," pp. 74–100, SIAM, Philadelphia, 1992.
13. D. Haussler, N. Littlestone, and M. K. Warmuth, Predicting $\{0, 1\}$-functions on randomly drawn points, *in* "Proceedings of the First Annual Workshop on Computational Learning Theory," pp. 280–296, Morgan Kaufmann, San Mateo, CA, 1988.
14. D. Haussler, M. Kearns, and R. Schapire, Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension, *in* "Proceedings of the Fourth Annual Workshop on Computational Learning Theory," pp. 61–74, Morgan Kaufmann, San Mateo, CA, 1991.
15. D. Haussler, J. Kivinen, and M. Warmuth, Tight worst-case loss bounds for predicting with expert advice, *in* "Computational Learning Theory: Second European Conference, EuroCOLT'95," pp. 69–83, Springer-Verlag, New York/Berlin, 1995.
16. N. Hernndorf, Best Φ-and N_ϕ-approximants in Orlicz spaces of vector valued functions, *Z. Wahrsch. Verw. Gebiete* **58** (1981), 309–329.
17. M. Kearns and R. Schapire, Efficient distribution-free learning of probabilistic concepts, *J. Comput. System Sci.* **48** (1994), 464–497.
18. J. Kivinen and M. Warmuth, Using experts for predicting continuous outcomes, *in* "Computational Learning Theory: EuroCOLT'93," pp. 109–120, Oxford Univ. Press, Oxford, 1994.
19. S. Kullback, A lower bound for discrimination in terms of variation, *IEEE Trans. Inform. Theory* **IT-13** (1967), 126–127.
20. L. LeCam, On some asymptotic properties of maximum likelihood estimates and related Bayes estimates, *Univ. California Publ. Stat.* **1** (1953), 277–330.
21. L. LeCam, On the asymptotics used to prove asymptotic normality of maximum likelihood estimates, *Ann. Math. Statist.* **41** (1970), 802–828.
22. N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear threshold algorithm, *Mach. Learning* **2** (1988), 258–318.
23. N. Littlestone and M. K. Warmuth, The weighted majority algorithm, *Inform. and Comput.* **108** (1994), 212–261.
24. N. Merhav and M. Feder, Universal sequential learning and decision from individual data sequence, *in* "Proceedings of the Fifth ACM Conference on Computational Learning Theory," pp. 413–427, Assoc. Comput. Mach., New York, 1992.
25. J. Rissanen, A universal prior for integers and estimation by minimum description length, *Ann. Statist.* **11** (1983), 416–431.
26. J. Rissanen, A universal data compression system, *IEEE Trans. Inform. Theory* **IT-29** (1983), 656–664.
27. J. Rissanen, Universal coding, information, prediction, and estimation, *IEEE Trans. Inform. Theory* **IT-30** (1984), 629–636.
28. J. Rissanen, Minimum description length principle, *in* "Encyclopedia of Statistical Sciences" (S. Kotz and N. L. Johnson, Eds.), Vol. 5, pp. 523–527, Wiley, New York, 1985.
29. J. Rissanen, Stochastic complexity, *J. R. Stat. Soc. B* **49** (1987), 223–239.
30. J. Rissanen, "Stochastic Complexity in Statistical Inquiry," Series in Computer Science, Vol. 15, World Scientific, Singapore, 1989.
31. K. Takeuchi, "Asymptotic Theory of Statistical Estimation," Kyooiku, Tokyo, 1974. [in Japanese]
32. V. G. Vovk, Aggregating strategies, *in* "Proceedings of the Third Annual Workshop on Computational Learning Theory," pp. 371–386, Morgan Kaufmann, 1990.
33. V. G. Vovk, A game of prediction with expert advice, *in* "Proceedings of the Eight Annual ACM Conference on Computational Learning Theory, 1995," pp. 51–60,.
34. K. Yaminishi, A loss bound model for on-line stochastic prediction strategies, *in* "Proceedings of the Fourth Annual Workshop on Computational Learning Theory," pp. 290–302, Morgan Kaufmann, San Mateo, CA, 1991.
35. K. Yamanishi, A learning criterion for stochastic rules, *Mach. Learning* **9** (1992), 165–203.
36. K. Yamanishi, On-line maximum likelihood prediction with respect to general loss functions, *in* "Computational Learning Theory: Second European Conference, EuroCOLT'95," pp. 84–98, Springer-Verlag, New York/Berlin, 1995.
37. K. Yamanishi, A loss bound model for on-line stochastic prediction algorithms, *Inform. and Comput.* **119**, No. 1 (1995), 39–54.